Statistics for Social Research III

Derrick Snow

Omega Graduate School

March 9, 2025

Professor

Dr. Taladay

***Assignment #3 – Essay***

1. Write a 5-page paper processing the development of an in-depth data analysis plan.

Include a title page, well-developed introduction and conclusion paragraphs, a references page, and in-text APA-formatted citations to support your responses.

**Data Analysis for Examining the Relationship between Land Use and Illness**

An interest of mine in sociology, among many, is the sociology of health and illness as well as geo-sociology. The research project I plan to carry out for my dissertation research combines both of those sociological interests, in examining the relationship between land use and illness. Moreover, what are the illnesses that have a correlation with certain land use? This research will not be about the why certain illnesses are associated with certain land use, but will identify if there are relationships and the strength of the relationship. The importance of this research is at least two fold, where different land use have different environmental impacts on the given area (Zaldo-Aubanell, et al., 2021), and moreover, land use is a reflection of the social milieu, of a given area, what people do for a living, their lifestyle and recreational activities, all that have an impact on a community’s health and illness.

From conducting a literature review, I find that there is a gap in research in taking a macro-sociological approach to understanding the relationship between illness and land use. There has been more micro-level research analyzing a particular illness/disease with a particular land use/land cover in a given geographical area. This research is not a small undertaking, but it is needed to further understand the relationship between health and illness and geography. The ancient Greek physician, Hippocrates, informed the world that there is a connection between geography and health, “If you want to learn about the health of a population, look at the air they breathe, the water they drink, and the places where they live” (Gracia & Koh, 2011, p.14). In modern healthcare research, known as GeoHealth, (Blanford, 2024), it has been found that a significant portion of health and illness is determined by zip codes or where we live, which has been found to be largely related to economic outcomes such as poverty or wealth and other socioeconomic variables (Embury et al , 2022), economic development, and localities of hospitals and health centers. This understanding gets to a proportion of the social drivers of health/social determinants of health, but are missing out on other proportions of social drivers of health, such as lifestyle of people within the geographical areas, which I believe land use analysis will be able to get at. Using the medical sociology framework of Health Lifestyle Theory brought forth by the medical sociologist William C. Cockerham, derived from the French sociologist Pierre Bourdieu’s concepts of fields and habitus, in my research of the relationship between land use and illness, will add to the understanding of social drivers of health and illness.

**Data**

There are two components to this research, one is the geographical or spatial, land use/land cover (LULC), spatial analysis is the analysis of data collected in a geographic area (Kent & Mardia, 2022). Land use, the purpose the land is utilized (Nedd et al., 2021), relates to land coverage the “specific landscape patterns and characteristics” (Nedd et al., 2021, p.2), such as; Developed High Intensity, Evergreen Forest, Grassland/Herbaceous, and Cultivated Crops to name a few. Land cover geographic data is derived from Landsat data, spatial data from satellite imaging. This data is ingested into geographic analysis system (GIS) software (Biu, et al, 2024) such an ArcGIS, as spatial features or vectors (Akindahunsi et al., 2024). The other component, non-spatial data or attribute data, as in illness/disease data, such as; hypertension, overweight and obesity, and depression to name a few, based on diagnosis information or International Classification of Diseases (ICD-10) from Electronic Health Records (EHR) systems (Zaldo-Aubanell, et al., 2021) or vital statistics data such as death records. The attribute data will need a geographic connector to analyze spatial and non-spatial data, such as zip codes of the patients which are commonly available in EHR and vital statistics data. So then for this research the illness/disease data will be connected geographically based on zip code, and the land use/land cover will need to be categorized by the majority land cover within a zip code area, or rather a zip code tabulation area (ZCTA), which estimates the zip code coverage area that is a geographic boundary. This will allow the research to be conducted based on a standard classification of areas for land use and illness.

The land use/land cover data has twenty different classifications and is at the interval-ratio level of measurement. The illness/disease data is at the interval-ratio level of measurement based on a rate and there will be ten to twenty categories of illness/disease. Thus, it will be a challenge to be able to analyze all the classifications and categories. However, given the task of the sociological researcher is to figure out how to measure social phenomena, this is exactly what I need to figure out, how to best measure the relationship between certain land use areas and certain illnesses. Given availability of the data in the archival/secondary electronic form, this research should be able to be conducted, it will take much work, but I think it is worthy of the work in order to add to the body of knowledge surrounding the importance of social drivers of health to the health and wellbeing of communities.

**Analysis**

Given that this research is a spatial analysis, extra spatial dynamics must be considered in statistical analysis, such as spatial autocorrelation, where nearby areas tend to be highly correlated with each other (Kent & Mardia, 2022), analyzed by the Global Moran’s I statistic. In my literature review of somewhat similar research, they were utilizing some type of regression, either linear or logistic regression. This makes sense where regression is a correlational modeling that is intended to model the outcome with the response variables in order to predict a given outcome. Given that there could be in the land use and illness to be measueed, with nominal and interval-ratio variables, logistic regression would be the model to utilize (Beltrán et al., 2021). There were some specific spatial regressions used in some of the research, such as Geographic Weighted Regression (GWR) (Tokey, 2021) and Land Use Regression (LUR) (Das, et al, 2024), I will further look into those models, but for now I will be using linear regression to examine the relationship between land cover/land use classifications and the rate of illness/disease within SPSS.

The land use/land cover by ZCTA data will consist of a column 20 land cover based on the proportion ZCTA measured of that given land cover. Then there would be a column for the value of the illness, for example.

A screenshot of a computer

AI-generated content may be incorrect.

This will allow me to do scatter plots, correlation tables, and linear regression models with each of the illness measures and all of the land cover categories. I will examine each model and will report on each and identify if certain land use had a relationship or not with each illness measure. There will be certain outputs to examine the models, such as correlations, the model summary of r and adjusted r2, and the ANOVA table, the F statistic andsignificant value. And the coefficient table with the significant values and the unstandardized B coefficients for each variable in the model.

This study will test data for normality and relevant assumptions of appropriate statistical procedures. for parametric procedures. This study will utilize Pearson’s Product Moment of Correlation (parametric) to test the hypotheses for statistically significant relationships. This study will include post-hoc statistical procedures of Bonferroni correction and Tukey's HSDaid the interpretation of the results.

**Conclusion**

For this study, it will be heavily reliant on the computing power of ArcGIS and statistical programs given the large amount of data that will be processed for this research. One of the benefits of utilizing archival/secondary data, is the amount of data that could be analyzed. However, caution needs to be taken, to stratify data as needed to gain an appropriate interpretation. I believe, I will need to stratify the analysis by regions, such as the Northwest, Southwest, Midwest, Northeast and Southeast of the United States. My assumption is that there could be regional differences with similar land cover.

I look forward to conducting this research as my dissertation topic, where it is a robust research project that can contribute to a larger body of health research. I am thankful to my professors for allowing me to pursue a research project that utilizes advanced social statistics.

**References**

Akindahunsi, T., Olulaja, O., Ajayi, O., Onyenegecha, I. P., Hanson, U., & Fadojutimi, B. (2024). Analytical tools in diseases epidemiology and surveillance: A review of literature. *International Journal of Applied Research*, *10*(9), 155–161. <https://doi.org/10.22271/allresearch.2024.v10.i9c.12018>

Beltrán, S., Arenas, D. J., López-Hinojosa, I. J., Tung, E. L., & Cronholm, P. F. (2021). Associations of race, insurance, and zip code-level income with nonadherence diagnoses in primary and specialty diabetes care. *The Journal of the American Board of Family Medicine*, *34*(5), 891–897. <https://doi.org/10.3122/jabfm.2021.05.200639>

Blanford, J. (2024). *Geographic information, geospatial technologies and spatial data science for health*. CRC Press.

Das, K., Das Chatterjee, N., Jana, D., & Bhattacharya, R. K. (2023). Application of land-use regression model with regularization algorithm to assess PM2.5 and PM10 concentration and health risk in Kolkata metropolitan. *Urban Climate*, *49*, 101473. <https://doi.org/10.1016/j.uclim.2023.101473>

Embury, J., Tsou, M.-H., Nara, A., & Oren, E. (2022). A Spatio-demographic perspective on the role of social determinants of health and chronic disease in determining a population’s vulnerability to COVID-19. *Preventing Chronic Disease*, *19*, 210414. <https://doi.org/10.5888/pcd19.210414>

Gracia, J.N., Koh, H.K. (2011). Promoting environmental justice. *American Journal of Public Health.* Suppl 1(Suppl 1):S14-6. doi: 10.2105/AJPH.2011.300406. Epub 2011 Oct 25. PMID: 22028454; PMCID: PMC3222484.

Kent, J. T., & Mardia, K. V. (2022). *Spatial analysis* (Vol. 72). John Wiley & Sons.

Nedd, R., Light, K., Owens, M., James, N., Johnson, E., & Anandhi, A. (2021). A Synthesis of land use/land cover studies: Definitions, classification systems, meta-studies, challenges and knowledge gaps on a global landscape. *Land*, *10*(9), 994. <https://doi.org/10.3390/land10090994>

Preye Winston Biu, Chinedu Nnamdi Nwasike, Olawe Alaba Tula, Chinedu Alex Ezeigweneme, & Joachim Osheyor Gidiagba. (2024). A review of GIS applications in public health surveillance. *World Journal of Advanced Research and Reviews*, *21*(1), 030–039. <https://doi.org/10.30574/wjarr.2024.21.1.2684>

Tokey, A. I. (2021). Spatial association of mobility and COVID-19 infection rate in the USA: A county-level study using mobile phone location data. *Journal of Transport & Health*, *22*, 101135. <https://doi.org/10.1016/j.jth.2021.101135>

Zaldo-Aubanell, Q., Serra, I., Sardanyés, J., Alsedà, L., & Maneja, R. (2021). Reviewing the reliability of land use and land cover data in studies relating human health to the environment. *Environmental Research*, *194*, 110578. <https://doi.org/10.1016/j.envres.2020.110578>